Alternative Approaches to Metadata Evaluation

dealing with

EDIT's, FILLER's, IP's and SU's

md-eval

Treats metadata as metadata events

rteval

Treats metadata as word annotations

The Performance Measures

md-eval

Word Coverage Error

- Applies to EWD and FWD

$$\text{Error} = \frac{\#\text{ ref DEPOD tokens not covered by sys DEPODs} + \#\text{ ref non-DEPOD tokens covered by sys DEPODs}}{\#\text{ ref DEPOD tokens}}$$

Boundary Error

- Applies to IPD and SUBD

$$\text{Error} = \frac{\#\text{ missed boundary tokens} + \#\text{ false alarm boundary tokens}}{\#\text{ ref boundary tokens}}$$

rteval

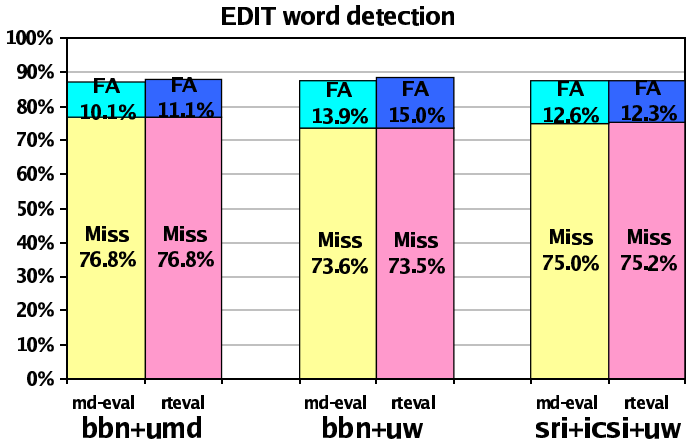
Slot Error

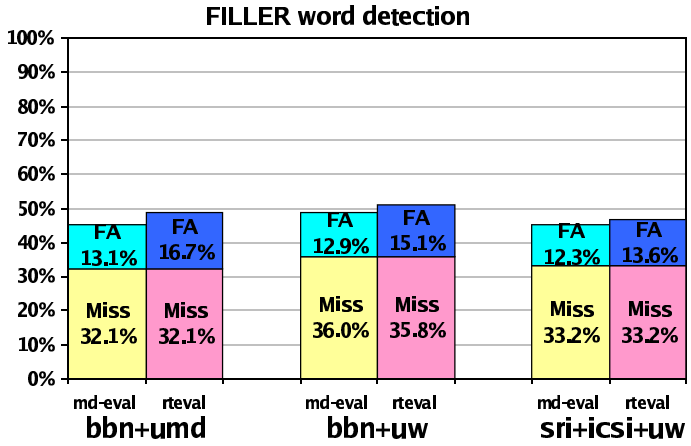
- Applies to EWD, FWD, IPD and SUBD

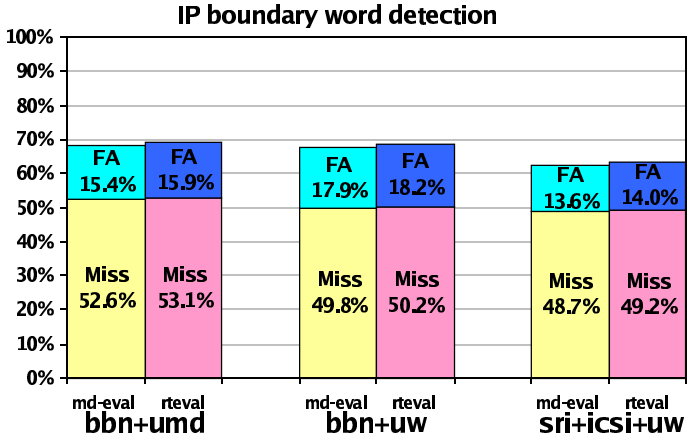
$$\text{Error} = \frac{\#\text{ sys \$TASK tokens that fail to align to ref \$TASK tokens} + \#\text{ ref \$TASK tokens that fail to align to sys \$TASK tokens}}{\#\text{ ref tokens with an active slot}}$$

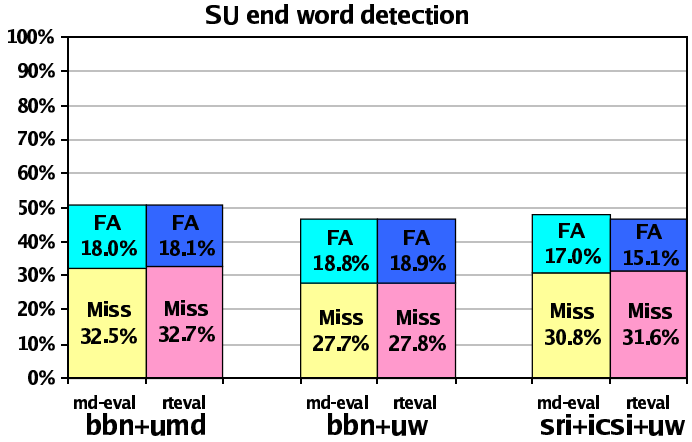
Comparison of Scores for CTS

md-eval *versus* rteval









Why do **md-eval** *and* **rteval** *yield different results?*

The primary reason

Different ways of counting errors

md-eval counts detection errors using the reference transcript as the basis for counting.

rteval counts detection errors using both the reference transcript and the system output words.

System output words classified as inserted during word alignment may contribute to the metadata word error count for **rteval**. System output words play no role in computing the word error count for **md-eval**.

Why do md-eval *and* rteval *yield different results?*

Secondary reasons

1. Different error weighting in word alignment

md-eval retains an STT-like (sclite) alignment optimization regarding filled pauses and fragments.

rteval uses equal weighting of all word token errors.

Equal weighting maximizes the flexibility of word alignments, so that (secondary) adjustments in word alignment are more likely to reduce the metadata word error rates.

Why do md-eval *and* rteval *yield different results?*

Secondary reasons

2. Different word alignment control strategies

md-eval constrains alignment to words that are temporally proximate (within one second of each other).

rteval constrains alignment to words that are in the same time segment (consistent with sclite).

These constraints produce different alignments, sometimes (dis)allowing words to match across segments boundaries, or separated by > 1 sec.

Why do md-eval *and* rteval *yield different results?*

Secondary reasons

3. Different handling of UEM exclusion zones

md-eval performs word alignment using *all* words, then counts errors only for those words that lie within the UEM evaluation intervals.

rteval discards words, *prior* to alignment, for all those words whose midpoints lie outside the UEM evaluation intervals, prior to alignment

Why do md-eval *and* rteval *yield different results?*

Secondary reasons

4. Promotion of lexical fp's to metadata events

md-eval accepts and processes reference and system output metadata without modification.

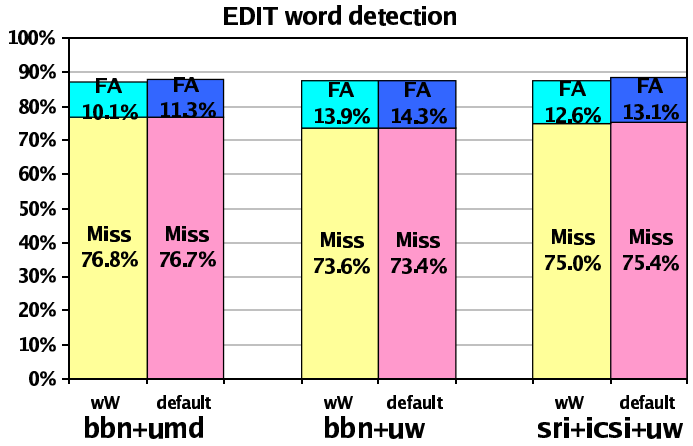
rteval creates metadata FILLER events when lexical "fp" tokens are encountered that are not subsumed within a FILLER metadata event.

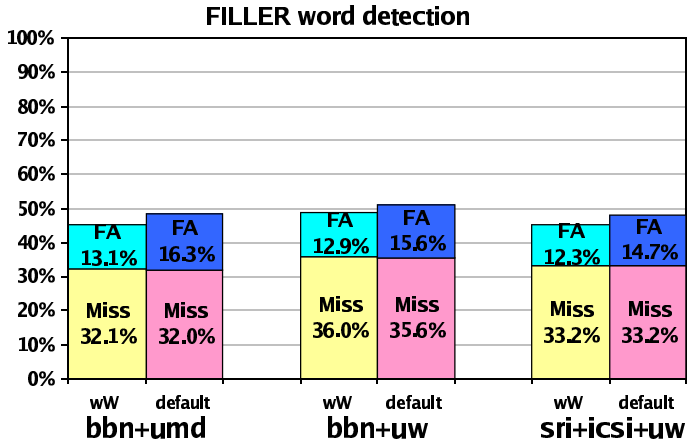
Major md-eval *parameters*

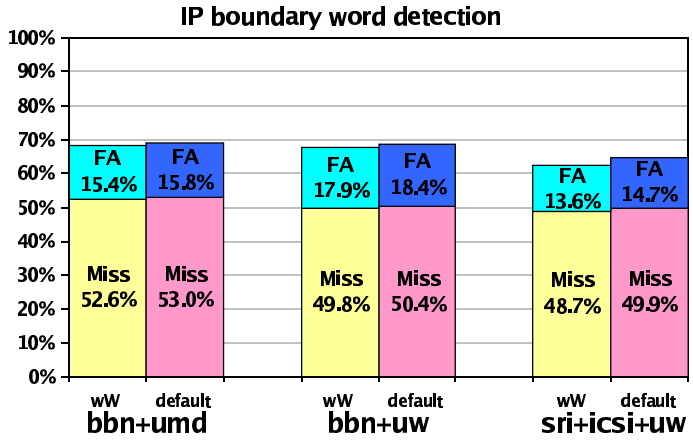
- **T**: Sets the maximum allowable time gap between system output metadata events and candidate reference metadata events. (default = 0.25 seconds)
- **W**: Changes metadata mapping so as to optimize metadata event overlap in terms of *words* rather than *time*.
- **w**: First performs (STT-like) word alignment and then modifies metadata times to agree with the resulting aligned word times.
- **t**: Sets the maximum allowable time gap between system output words and candidate reference cohorts. (default = 1.0 seconds)

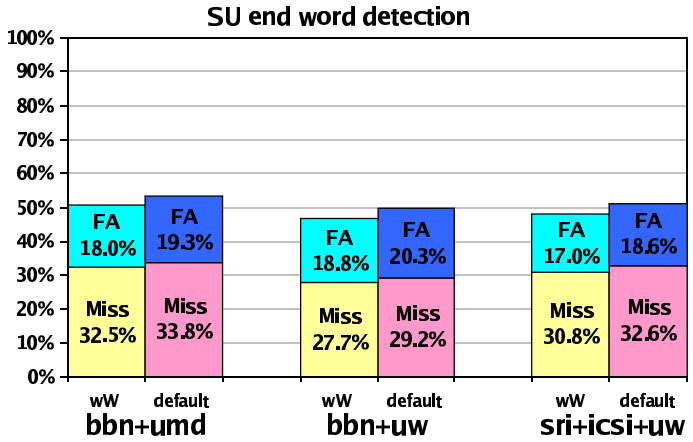
Comparison of md-eval *Scores for CTS*

md-eval (official) versus *md-eval (default)*







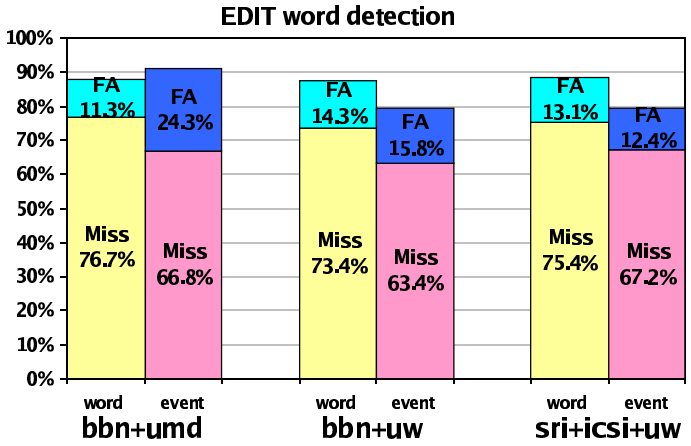


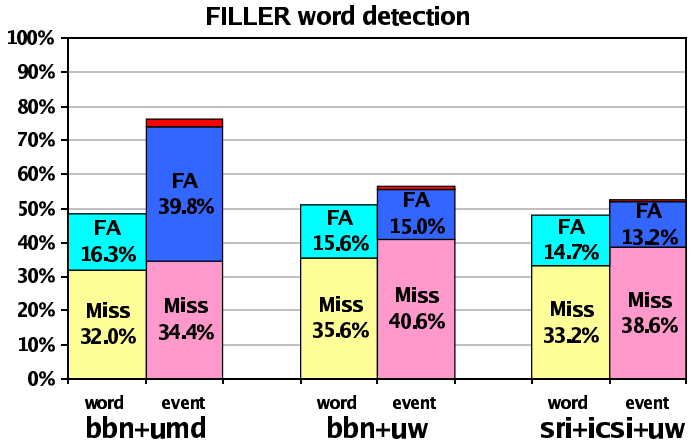
md-eval *Performance Measures*

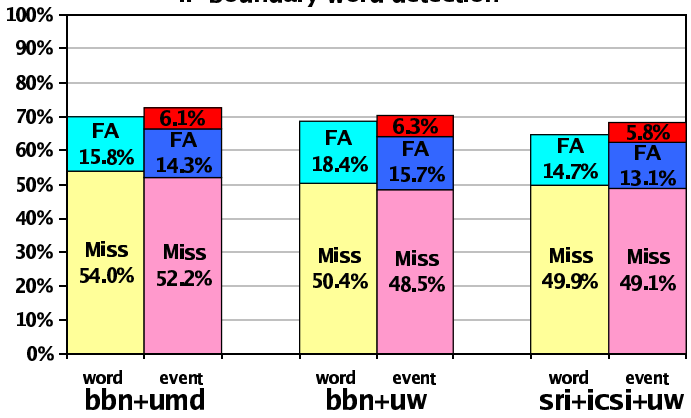
- Event Word Detection Errors (the official score)
  - Miss
  - False Alarm
- Event Detection Errors
  - Miss
  - False Alarm
  - Type Error
- Event Type Confusion Matrix
  (system output type versus reference type)
- Event Offset Histogram (for detected events)
  - For start point
  - For end point

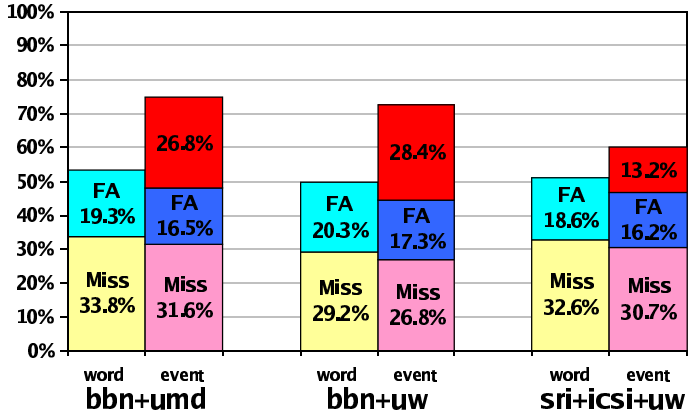
Comparison of md-eval Scores for CTS

Word Detection versus Event Detection



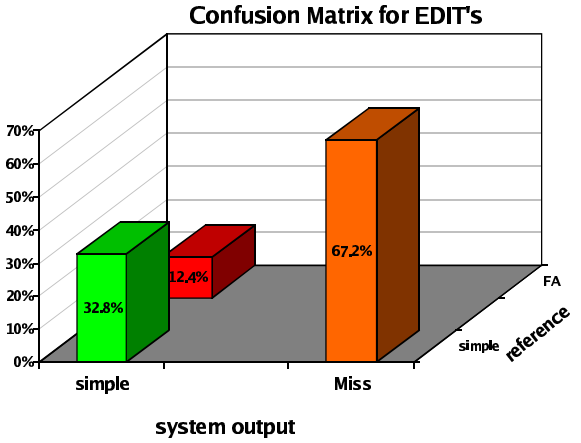


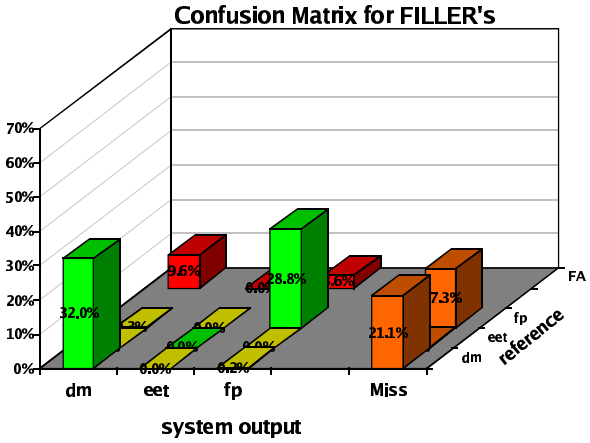


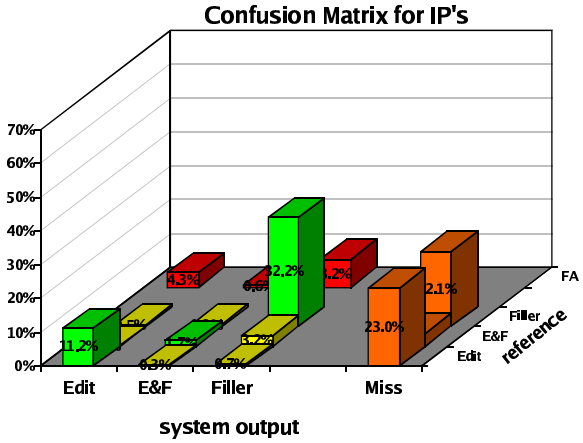


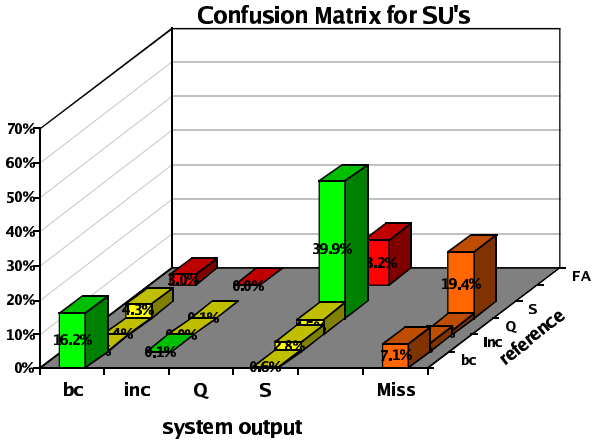
Event Type Confusion Matrices for CTS

(SRI+ICSI+UW results)









Event Offsets in Words for CTS

(SRI+ICSI+UW results)